

**What Is Claimed Is:**

1           1.       A method for characterizing a document with respect to clusters of  
2 conceptually related words, comprising:  
3           receiving the document, wherein the document contains a set of words;  
4           selecting candidate clusters of conceptually related words that are related  
5 to the set of words;  
6           wherein the candidate clusters are selected using a model that explains  
7 how sets of words are generated from clusters of conceptually related words; and  
8           constructing a set of components to characterize the document, wherein  
9 the set of components includes components for candidate clusters, wherein each  
10 component indicates a degree to which a corresponding candidate cluster is  
11 related to the set of words.

1           2.       The method of claim 1, wherein the model is a probabilistic model,  
2 which contains nodes representing random variables for words and for clusters of  
3 conceptually related words.

1           3.       The method of claim 2, wherein each component in the set of  
2 components indicates a degree to which a corresponding candidate cluster is  
3 active in generating the set of words.

1           4.       The method of claim 3,  
2 wherein nodes in the probabilistic model are coupled together by weighted  
3 links; and  
4 wherein if a cluster node in the probabilistic model fires, a weighted link  
5 from the cluster node to another node can cause the other node to fire.

1           5.       The method of claim 4, wherein if a node has multiple parent  
2 nodes that are active, the probability that the node does not fire is the product of  
3 the probabilities that links from the active parent nodes do not fire.

1           6.       The method of claim 2, wherein the probabilistic model includes a  
2 universal node that is always active and that has weighted links to all cluster  
3 nodes.

1           7.       The method of claim 4, wherein selecting the candidate clusters  
2 involves:  
3           constructing an evidence tree by starting with terminal nodes associated  
4 with the set of words in the document, and following links in the reverse direction  
5 to parent cluster nodes;  
6           using the evidence tree to estimate a likelihood that each parent cluster  
7 node was active in generating the set of words; and  
8           selecting a parent cluster node to be a candidate cluster node based on its  
9 estimated likelihood.

1           8.       The method of claim 7, wherein estimating the likelihood that a  
2 given parent node is active in generating the set of words may involve  
3 considering:  
4           the unconditional probability that the given parent node is active;  
5           conditional probabilities that the given parent node is active assuming  
6 parent nodes of the given parent node are active; and  
7           conditional probabilities that the given parent node is active assuming  
8 child nodes of the given parent node are active.

1           9.     The method of claim 8, wherein considering the conditional  
2 probabilities involves considering weights on links between nodes.

1           10.    The method of claim 7 wherein estimating the likelihood that a  
2 given parent node is active in generating the set of words involves marking  
3 terminal nodes during the estimation process to ensure that terminal nodes are not  
4 factored into the estimation more than once.

1           11.    The method of claim 7, wherein constructing the evidence tree  
2 involves pruning unlikely nodes from the evidence tree.

1           12.    The method of claim 3, wherein during construction of the set of  
2 components, the degree to which a candidate cluster is active in generating the set  
3 of words is determined by calculating a probability that a candidate cluster is  
4 active in generating the set of words.

1           13.    The method of claim 3, wherein during construction of the set of  
2 components, the degree to which a candidate cluster is active in generating the set  
3 of words is determined by multiplying a probability that a candidate cluster is  
4 active in generating the set of words by an activation for the candidate cluster,  
5 wherein the activation indicates how many links from the candidate cluster to  
6 other nodes are likely to fire.

1           14.    The method of claim 1, wherein constructing the set of components  
2 involves normalizing the set of components.

1           15.     The method of claim 3, wherein constructing the set of components  
2 involves approximating a probability that a given candidate cluster is active over  
3 states of the probabilistic model that could have generated the set of words.

1           16.     The method of claim 15, wherein approximating the probability  
2 involves:  
3           selecting states for the probabilistic model that are likely to have generated  
4 the set of words in the document; and  
5           considering only selected states while calculating the probability that the  
6 given candidate cluster is active.

1           17.     The method of claim 16, wherein selecting a state that is likely to  
2 have generated the set of words involves:  
3           randomly selecting a starting state for the probabilistic model; and  
4           performing hill-climbing operations beginning at the starting state to reach  
5 a state that is likely to have generated the set of words.

1           18.     The method of claim 17, wherein performing the hill-climbing  
2 operations involves periodically changing states of individual candidate clusters  
3 without regards to an objective function for the hill-climbing operations to explore  
4 states of the probabilistic model that are otherwise unreachable through hill-  
5 climbing operations.

1           19.     The method of claim 18, wherein changing a state of an individual  
2 candidate cluster involves temporarily fixing the changed state to produce a local  
3 optimum for the objective function, which includes the changed state.

1           20.     The method of claim 1, wherein the document can include:  
2           a web page; or  
3           a set of terms from a query.

1           21.     A computer-readable storage medium storing instructions that  
2     when executed by a computer cause the computer to perform a method for  
3     characterizing a document with respect to clusters of conceptually related words,  
4     the method comprising:  
5           receiving the document, wherein the document contains a set of words;  
6           selecting candidate clusters of conceptually related words that are related  
7     to the set of words;  
8           wherein the candidate clusters are selected using a model that explains  
9     how sets of words are generated from clusters of conceptually related words; and  
10          constructing a set of components to characterize the document, wherein  
11     the set of components includes components for candidate clusters, wherein each  
12     component indicates a degree to which a corresponding candidate cluster is  
13     related to the set of words.

1           22.     The computer-readable storage medium of claim 21, wherein the  
2     model is a probabilistic model, which contains nodes representing random  
3     variables for words and for clusters of conceptually related words.

1           23.     The computer-readable storage medium of claim 22, wherein each  
2     component in the set of components indicates a degree to which a corresponding  
3     candidate cluster is active in generating the set of words.

1           24.     The computer-readable storage medium of claim 23,

2 wherein nodes in the probabilistic model are coupled together by weighted  
3 links; and

4 wherein if a cluster node in the probabilistic model fires, a weighted link  
5 from the cluster node to another node can cause the other node to fire.

1 25. The computer-readable storage medium of claim 24, wherein if a  
2 node has multiple parent nodes that are active, the probability that the node does  
3 not fire is the product of the probabilities that links from the active parent nodes  
4 do not fire.

1 26. The computer-readable storage medium of claim 22, wherein the  
2 probabilistic model includes a universal node that is always active and that has  
3 weighted links to all cluster nodes.

1 27. The computer-readable storage medium of claim 24, wherein  
2 selecting the candidate clusters involves:  
3 constructing an evidence tree by starting with terminal nodes associated  
4 with the set of words in the document, and following links in the reverse direction  
5 to parent cluster nodes;  
6 using the evidence tree to estimate a likelihood that each parent cluster  
7 node was active in generating the set of words; and  
8 selecting a parent cluster node to be a candidate cluster node based on its  
9 estimated likelihood.

1 28. The computer-readable storage medium of claim 27, wherein  
2 estimating the likelihood that a given parent node is active in generating the set of  
3 words may involve considering:

4 the unconditional probability that the given parent node is active;  
5 conditional probabilities that the given parent node is active assuming  
6 parent nodes of the given parent node are active; and  
7 conditional probabilities that the given parent node is active assuming  
8 child nodes of the given parent node are active.

1 29. The computer-readable storage medium of claim 28, wherein  
2 considering the conditional probabilities involves considering weights on links  
3 between nodes.

1 30. The computer-readable storage medium of claim 27, wherein  
2 estimating the likelihood that a given parent node is active involves marking  
3 terminal nodes during the estimation process to ensure that terminal nodes are not  
4 factored into the estimation more than once.

1 31. The computer-readable storage medium of claim 27, wherein  
2 constructing the evidence tree involves pruning unlikely nodes from the evidence  
3 tree.

1 32. The computer-readable storage medium of claim 23, wherein  
2 during construction of the set of components, the degree to which a candidate  
3 cluster is active in generating the set of words is determined by calculating a  
4 probability that a candidate cluster is active in generating the set of words.

1 33. The computer-readable storage medium of claim 23, wherein  
2 during construction of the set of components, the degree to which a candidate  
3 cluster is active in generating the set of words is determined by multiplying a

4 probability that a candidate cluster is active in generating the set of words by an  
5 activation for the candidate cluster, wherein the activation indicates how many  
6 links from the candidate cluster to other nodes are likely to fire.

1 34. The computer-readable storage medium of claim 21, wherein  
2 constructing the set of components involves normalizing the set of components.

1 35. The computer-readable storage medium of claim 23, wherein  
2 constructing the set of components involves approximating a probability that a  
3 given candidate cluster is active over states of the probabilistic model that could  
4 have generated the set of words.

1 36. The computer-readable storage medium of claim 35, wherein  
2 approximating the probability involves:  
3 selecting states for the probabilistic model that are likely to have generated  
4 the set of words in the document; and  
5 considering only selected states while calculating the probability that the  
6 given candidate cluster is active.

1 37. The computer-readable storage medium of claim 36, wherein  
2 selecting a state that is likely to have generated the set of words involves:  
3 randomly selecting a starting state for the probabilistic model; and  
4 performing hill-climbing operations beginning at the starting state to reach  
5 a state that is likely to have generated the set of words.

1 38. The computer-readable storage medium of claim 37, wherein  
2 performing the hill-climbing operations involves periodically changing states of



3 individual candidate clusters without regards to an objective function for the hill-  
4 climbing operations to explore states of the probabilistic model that are otherwise  
5 unreachable through hill-climbing operations.

1 39. The computer-readable storage medium of claim 38, wherein  
2 changing a state of an individual candidate cluster involves temporarily fixing the  
3 changed state to produce a local optimum for the objective function, which  
4 includes the changed state.

1 40. The computer-readable storage medium of claim 21, wherein the  
2 document can include:  
3 a web page; or  
4 a set of terms from a query.

1 41. An apparatus for characterizing a document with respect to clusters  
2 of conceptually related words, comprising:  
3 a receiving mechanism, configured to receive the document, wherein the  
4 document contains a set of words;  
5 a selection mechanism configured to select candidate clusters of  
6 conceptually related words that are related to the set of words;  
7 wherein the candidate clusters are selected using a model that explains  
8 how sets of words are generated from clusters of conceptually related words; and  
9 a component construction mechanism configured to construct a set of  
10 components to characterize the document, wherein the set of components includes  
11 components for candidate clusters, wherein each component indicates a degree to  
12 which a corresponding candidate cluster is related to the set of words.

1           42.     The apparatus of claim 41, wherein the model is a probabilistic  
2 model, which contains nodes representing random variables for words and for  
3 clusters of conceptually related words.

1           43.     The apparatus of claim 42, wherein each component in the set of  
2 components indicates a degree to which a corresponding candidate cluster is  
3 active in generating the set of words.

1           44.     The apparatus of claim 43,  
2 wherein nodes in the probabilistic model are coupled together by weighted  
3 links; and  
4 wherein if a cluster node in the probabilistic model fires, a weighted link  
5 from the cluster node to another node can cause the other node to fire.

1           45.     The apparatus of claim 44, wherein if a node has multiple parent  
2 nodes that are active, the probability that the node does not fire is the product of  
3 the probabilities that links from the active parent nodes do not fire.

1           46.     The apparatus of claim 43, wherein the probabilistic model  
2 includes a universal node that is always active and that has weighted links to all  
3 cluster nodes.

1           47.     The apparatus of claim 44, wherein the selection mechanism is  
2 configured to:  
3 construct an evidence tree by starting with terminal nodes associated with  
4 the set of words in the document, and following links in the reverse direction to  
5 parent cluster nodes;

6           use the evidence tree to estimate a likelihood that each parent cluster node  
7   was active in generating the set of words; and to  
8           select a parent cluster node to be a candidate cluster node based on its  
9   estimated likelihood.

1           48.    The apparatus of claim 47, wherein while estimating the likelihood  
2   that a given parent node is active in generating the set of words, the selection  
3   mechanism is configured to consider at least one of the following:  
4           the unconditional probability that the given parent node is active;  
5           conditional probabilities that the given parent node is active assuming  
6   parent nodes of the given parent node are active; and  
7           conditional probabilities that the given parent node is active assuming  
8   child nodes of the given parent node are active.

1           49.    The apparatus of claim 48, wherein while considering the  
2   conditional probabilities, the selection mechanism is configured to consider  
3   weights on links between nodes.

1           50.    The apparatus of claim 47, wherein while estimating the likelihood  
2   that a given parent node is active in generating the set of words, the selection  
3   mechanism is configured to mark terminal nodes during the estimation process to  
4   ensure that terminal nodes are not factored into the estimation more than once.

1           51.    The apparatus of claim 47, wherein while constructing the  
2   evidence tree, the selection mechanism is configured to prune unlikely nodes from  
3   the evidence tree.

1           52.     The apparatus of claim 43, wherein while constructing a given  
2 component in the set of components, the component construction mechanism is  
3 configured to determine the degree to which a candidate cluster is active in  
4 generating the set of words by calculating a probability that a candidate cluster is  
5 active in generating the set of words.

1           53.     The apparatus of claim 43, wherein while constructing a given  
2 component in the set of components, the component construction mechanism is  
3 configured to determine the degree to which a candidate cluster is active in  
4 generating the set of words by multiplying a probability that a candidate cluster is  
5 active in generating the set of words by an activation for the candidate cluster,  
6 wherein the activation indicates how many links from the candidate cluster to  
7 other nodes are likely to fire.

1           54.     The apparatus of claim 41, wherein the component construction  
2 mechanism is configured to normalize the set of components.

1           55.     The apparatus of claim 43, wherein the component construction  
2 mechanism is configured to approximate a probability that a given candidate  
3 cluster is active over states of the probabilistic model that could have generated  
4 the set of words.

1           56.     The apparatus of claim 55, wherein while approximating the  
2 probability, the component construction mechanism is configured to:  
3           select states for the probabilistic model that are likely to have generated  
4 the set of words in the document; and to

5           consider only selected states while calculating the probability that the  
6   given candidate cluster is active.

1           57.    The apparatus of claim 56, wherein while selecting a state that is  
2   likely to have generated the set of words, the component construction mechanism  
3   is configured to:

4           randomly select a starting state for the probabilistic model; and to  
5           perform hill-climbing operations beginning at the starting state to reach a  
6   state that is likely to have generated the set of words.

1           58.    The apparatus of claim 58, wherein while performing the hill-  
2   climbing operations, the component construction mechanism is configured to  
3   periodically change states of individual candidate clusters without regards to an  
4   objective function for the hill-climbing operations to explore states of the  
5   probabilistic model that are otherwise unreachable through hill-climbing  
6   operations.

1           59.    The apparatus of claim 58, wherein while changing a state of an  
2   individual candidate cluster, the component construction mechanism is configured  
3   to temporarily fix the changed state to produce a local optimum for the objective  
4   function, which includes the changed state.

1           60.    The apparatus of claim 41, wherein the document can include:  
2           a web page; or  
3           a set of terms from a query.

1           61.     A computer-readable storage medium containing a data structure  
2     that facilitates characterizing a document with respect to clusters of conceptually  
3     related words, the data structure comprising:  
4           a probabilistic model that contains nodes representing random variables  
5     for words and for clusters of conceptually related words;  
6           wherein nodes in the probabilistic model are coupled together by weighted  
7     links;  
8           wherein if a cluster node in the probabilistic model fires, a weighted link  
9     from the cluster node to another node can cause the other node to fire; and  
10          wherein the other code can be associated with a word or a cluster.

1           62.     The computer-readable storage medium of claim 61, wherein the  
2     probabilistic model includes a universal node that is always active and that has  
3     weighted links to all cluster nodes.